

GSE71831 README

This README describes the various files and file formats that have been uploaded to GEO accession GSE71831 as part of the study “Deletion of DXZ4 on the human inactive X chromosome eliminates superdomains and impairs gene silencing” (Darrow & Huntley, et al., PNAS 2016).

For this study, we performed 20 in situ Hi-C experiments on five different cell lines (RPE1- Δ DXZ4a, RPE1- Δ DXZ4i, RPE1-WT, Patski, and AG08312). We performed one COLA experiment on GM12878. We performed 6 RNA-seq experiments (on RPE1- Δ DXZ4i and RPE1-WT).

As part of this study, we have uploaded data at various processing stages, including:

- raw sequence data from RNA-seq experiments (FASTQ format)
- transcript abundances from RNA-seq
- raw sequence data from Hi-C experiments (FASTQ format)
- Hi-C data post-alignment and duplicate filtering (merged_nodups.txt format)
- Hi-C collisions data post (modified sam format)
- binary Hi-C files (.hic format)
- contact matrices and associated normalization and 1-D expected vectors (tar archives)
- loop annotations and contact domain annotations

The various files and file formats associated with each processing stage will be described in more detail below.

RNA-Seq Data:

Accession Titles: DarrowHuntley-2016-RNA00*.

Cell Types: RPE1- Δ DXZ4i and RPE1-WT

Data from the six RNA-seq experiments is provided, with each experiment labeled as “DarrowHuntley-2016-RNA” followed by a three digit number (e.g., DarrowHuntley-2016-RNA001). RNA001-RNA003 correspond to RPE1-WT cells, and RNA004-RNA006 correspond to RPE1- Δ DXZ4i. FASTQ data for each experiment is available through SRA under the Sample Record for each experiment. Similarly, transcript abundances as calculated by kallisto are provided for the maternal and paternal X chromosome alleles under the Sample Record for each experiment.

Hi-C Raw Data:

Accession Titles: DarrowHuntley-2016-HIC*.

Cell types: RPE1- Δ DXZ4a, RPE1- Δ DXZ4i, RPE1-WT, Patski, AG08312, and GM12878

FASTQ data for all the 21 Hi-C and COLA experiments performed in this study are available through SRA under the Sample Record for each experiment. Each experiment is labeled with a unique identifier: “DarrowHuntley-2016-HIC” followed by a three digit number (for example, HIC001). Full details for each experiment (including cell type, number of reads sequenced, number of duplicates, etc.) are available in Supplemental Table S1 of Darrow & Huntley, et al., PNAS, 2016.

Hi-C Raw Data for Previous Libraries:

Accession Titles: RaoHuntley-2014-HIC*.

Cell types: GM12878, IMR90, HMEC, NHEK, K562

In addition to the 21 new libraries created for this study, we also have links to an additional 74 Hi-C libraries that were created for our previous study (Rao & Huntley, et al., Cell 2014) and which we used for our analysis of triples. Each experiment is labeled with a unique identifier: "RaoHuntley-2014-HIC" followed by a three digit number (for example, HIC001). Full details for each experiment (including cell type, number of reads sequenced, number of duplicates, etc.) are available in Supplemental Table S1 of Rao & Huntley, et al., Cell, 2014.

Hi-C Collisions Data:

Filenames: GSM*_RaoHuntley-2014-HIC*.collisions.txt.gz,
GSM1847540_DarrowHuntley-2015-HIC021.collisions.txt.gz

Cell types: GM12878, IMR90, HMEC, NHEK, K562

Under the sample accessions for the 74 libraries in Rao & Huntley, et al., Cell 2014 that we use, and for DarrowHuntley-2015-HIC21, the COLA library we created for this study, we provide collisions data. This data provides all the triples in the library, where a triple is defined as a single Hi-C contact aligning to three locations on the genome. This means that one of the paired reads aligned to a single location, and the other read aligned to two locations in the genome. Indicating the three alignment locations as A, B, and C, and the two reads as 1 and 2, the format of the file is a modified sam format:

```
read_name1 strandA chromosomeA positionA mapqA cigarA read1
read_name1_or_2 strandB chromosomeB positionB mapqB cigarB read1_or
_2
read_name2 strandC chromosomeC positionC mapqC cigarC read2
```

where:

read_name = the name of the read pair as seen in the FASTQ files. Can either be the read name of the first or second in the paired read.

strand = the strand that the read maps to (0=forward, 16=reverse)

chromosome = the chromosome that the read maps to: [1-22,X,Y,MT] for human (b37 reference genome), [1-20,X,Y] for rhesus macaque (rheMac2 reference genome), [chr1-chr19,chrX,chrY,chrM] for mouse (mm10 reference genome)

position = the position on the chromosome that the read maps to

mapq = the mapping quality score returned by BWA (see Section II.a.1 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014 for full details about the alignment process.)

cigar = the CIGAR string for the alignment

read = the read's sequence. Can either be the first or second of the two paired reads.

Post-alignment and Duplicate Filtered Data:

Filenames: GSM*_DarrowHuntley-2015-HIC*_merged_nodups.txt.gz

Cell types: RPE1-ΔDXZ4a, RPE1-ΔDXZ4i, RPE1-WT, Patski, AG08312, and GM12878

Under each Sample Record, there is a file available named GSM*_DarrowHuntley-2015-HIC*_merged_nodups.txt.gz, where GSM* is the GEO Sample ID and HIC* is the unique sample identifier from Supplemental Table S1 of Darrow & Huntley, et al., PNAS, 2016. This file contains post-alignment and post-duplicate filtering data for each HIC experiment.

At this stage of processing, read pairs where one or both ends do not align to the reference genome have already been removed, as well as chimeric ambiguous reads (see Section II.a.2 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014 for a definition of chimeric ambiguous reads). In addition, duplicate reads (reads where both ends align to within +/- 4bp of each other) have been removed as well (see Section II.a.3 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014 for a full description of duplicate removal). Full details of the Hi-C processing pipeline used in this study are provided in Section II.a. of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014.

Each line of the merged_nodups file represents a single Hi-C read pair that has passed the alignment and duplicate removal stages. The format of each line of the file is:

```
strand1 chromosome1 position1 fragmentindex1 strand2 chromosome2 position2  
fragmentindex2 mapq1 cigar1 seq1 mapq2 cigar2 seq2 read_name1 read_name2
```

where:

“1” or “2” suffix = 1st or 2nd read in a pair

strand = the strand that the read maps to (0=forward, 16=reverse)

chromosome = the chromosome that the read maps to: [1-22,X,Y,MT] for human (b37 reference genome), [1-20,X,Y] for rhesus macaque (rheMac2 reference genome), [chr1-chr19,chrX,chrY,chrM] for mouse (mm10 reference genome)

position = the position on the chromosome that the read maps to

fragmentindex = the index of the interval demarcated by restrictions sites in the genome, starting with 0 for the interval preceding the first restriction site

mapq = the mapping quality score returned by BWA (see Section II.a.1 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014 for full details about the alignment process.)

cigar = CIGAR string (SAM format)

seq = read sequence

read_name = the name of the read as seen in the FASTQ files

Binary (.hic) Files:

Filenames: GSE71831_[CELL_TYPE].hic.gz, GSE71831_[CELL_TYPE]_30.hic.gz, GSE71831_[CELL_TYPE]_maternal.hic.gz, GSE71831_[CELL_TYPE]_paternal.hic.gz, GSE71831_[CELL_TYPE]-chrX_diploid.hic.gz

Cell types: RPE1-ΔDXZ4a, RPE1-ΔDXZ4i, RPE1-WT, Patski, AG08312, and GM12878

Under the main series record (GSE71831), we provide binary .hic files for all of the maps, in .hic format. This is a file format that can be read by Juicebox, a visualization tool (<http://aidenlab.org/juicebox/>) for viewing contact maps. (They are also preloaded into the latest installment of Juicebox.) The .hic file format can also be read by Juicer

Command Line Tools (<http://aidenlab.org/commandlinetools/>), which can dump contact matrices and their normalization vectors, annotate the contact domains and loops and in the map, among other capabilities.

We provide two .hic files for each of the unphased datasets. Data from multiple libraries corresponding to the same cell line were pooled (see Supplemental Table S1 of Darrow & Huntley, et al., PNAS, 2016 to see which libraries correspond to which cell type). The .hic files are named as GSE71831_[CELL_TYPE].hic.gz and , GSE71831_[CELL_TYPE]_30.hic.gz. For example, for the Patski unphased data, we provide “Patski.hic” and “Patski_30.hic”. [CELL_TYPE].hic maps contain contact matrices constructed from all read pairs that uniquely align to the genome (MAPQ>0). [CELL_TYPE]_30.hic maps contain files associated with the contact matrices constructed from all read pairs that map to the genome with a MAPQ>=30. In general, we recommend using MAPQGE30 unless there is a good understanding of why MAPQG0 is necessary.

In addition to these .hic files, we provide phased .hic files for the cell lines that were phased. For information on how these maps were phased, see Supplemental Information of Darrow & Huntley, et al., PNAS, 2016.

For the phased data from the Patski cell line, we provide Patski_maternal.hic and Patski_paternal.hic. These maps contain data for all of the chromosomes of the allele type, with both intra- and inter-chromosomal contacts. For example, one can examine contacts between maternal chromosome 1 and maternal chromosome 2 using Patski_maternal.hic. Contacts between chromosomes of differing allele types (for example maternal chromosome 1 and paternal chromosome 2) is not available.

For each of the RPE1 lines (RPE1-WT, RPE1-deltaDXZ4a, RPE1-deltaDXZ4i), we provide the phased data for X chromosome only. There is only one .hic file associated with the phased data, per cell line, titled [CELL-TYPE]-chrX_diploid.hic. For example, the phased data from the RPE1 WT cell line can be found in RPE1-WT-chrX_diploid.hic. These hic files contain the maternal and paternal chrX contact data. Note that in this hic file there are only two chromosomes, and they are labeled as “chrXmat” and “chrXpat”, for the maternal and paternal X chromosomes. One can also access the contacts between the two alleles.

Contact matrices and Associated Normalization and Expected Vectors:

Filenames: GSE71831_[CELL_TYPE]_Interchromosomal.tar.gz,
GSE71831_[CELL_TYPE]_Intrachromosomal.tar.gz

Cell types: RPE1-ΔDXZ4a, RPE1-ΔDXZ4i, RPE1-WT, Patski, AG08312, and GM12878

Under the main series record (GSE71831), there are a set of tar archives that contain the raw observed intra-chromosomal and inter-chromosomal contact matrices for each cell type analyzed in our study as well as normalization vectors to transform the raw matrices into normalized matrices and 1-d expected vectors to transform the observed matrices into O/E matrices. For more information on how these 1-d vectors are computed, please see the Extended Experimental Procedures of Rao & Huntley, et al., Cell, 2014.

The intrachromosomal contact matrix archives are labeled GSE71831_[CELL_TYPE]_Intrachromosomal.tar.gz, where [CELL_TYPE] denotes the cell type (e.g. Patski), and which contain intrachromosomal contact matrices at 8 different base pair delimited resolutions (1 Mb, 500kb, 250kb, 100kb, 50kb, 25kb, 10kb, and 5kb). At each resolution, files associated with each chromosome are provided in a separate subdirectory.

Within each chromosome subdirectory, there are further subdirectories that specify the data type: "MAPQG0", "MAPQGE30", and, if available, "Phased". Both MAPQG0 and MAPQGE30 contained unphased data. The MAPQG0 subdirectory contains files associated with the contact matrices constructed from all read pairs that uniquely align to the genome (MAPQ>0). The MAPQGE30 subdirectory contains files associated with the contact matrices constructed from all read pairs that map to the genome with a MAPQ>=30. In general, we recommend using MAPQGE30 unless there is a good understanding of why MAPQG0 is necessary.

Phased data is additionally provided for all chromosomes of Patski, and for chrX of the three RPE1 cell lines. (For information on how these maps were phased, see Supplemental Information of Darrow & Huntley, et al., PNAS, 2016.) The contact maps and 1-d vectors, in identical format to the unphased data, is provided for maternal and paternal alleles in Phased/maternal and Phased/paternal respectively.

Intrachromosomal contact data can be found by navigating to the cell, resolution, chromosome, and type folder. For example, the files associated with the contact matrix for chromosome 1 at 5 kb resolution using MAPQ>=30 read pairs are in the subdirectory:

CELL_TYPE/5kb_resolution_intrachromosomal/chr1/MAPQGE30/

Similarly, the files associated with the contact matrix for chromosome X at 5 kb resolution on the maternal allele are in the subdirectory:

CELL_TYPE/5kb_resolution_intrachromosomal/chrX/Phased/maternal

In each intrachromosomal data folder, there are 7 text files associated with each intrachromosomal contact matrix ([CHR] represents the chromosome number and [RES] represents the resolution in kilobases):

chr[CHR]_[RES]kb.RAWobserved
chr[CHR]_[RES]kb.KRnorm
chr[CHR]_[RES]kb.VCnorm
chr[CHR]_[RES]kb.SQRTVCnorm
chr[CHR]_[RES]kb.RAWexpected
chr[CHR]_[RES]kb.KRexpected
chr[CHR]_[RES]kb.VCexpected
chr[CHR]_[RES]kb.SQRTVCexpected

(See the Glossary Appendix for explanations of these file types.) If a file is missing or empty, that means that there was not sufficient resolution to be able to compute the norm or expected vector.

*.RAWobserved is a text file with the raw observed contact matrix in sparse matrix notation. Each line has three fields: i, j, and $M_{i,j}$. (i and j are written as the left edge of the bin at a given resolution; for example, at 100 kb resolution, the entry corresponding to the first row and tenth column of the matrix would correspond to $M_{i,j}$, where $i=0$, $j=900000$). Only the upper triangle of the matrix is provided (i.e. $i \leq j$), the matrix is symmetric, so $M_{i,j} = M_{j,i}$.

The three *norm files are normalization vectors that can be used to transform the raw contact matrices M into normalized matrices M^* . (See section II.b of the Extended Experimental Procedures of Rao & Huntley, et al., Cell, 2014 for more information about the different types of normalizations.) Each file is ordered such that the first line of the normalization vector file is the norm factor for the first row/column of the corresponding raw contact matrix, the second line is the factor for the second row/column of the contact matrix, and so on. To normalize, an entry $M_{i,j}$ in a *.RAWobserved file, divide the entry by the corresponding norm factors for i and j.

For example, here is a line from the Rhesus-Macaque 1000kb chr1 MAPQGE30 raw observed contact matrix (/Rhesus-Macaque/1000kb_resolution_intrachromosomal/chr1/MQGE30/chrX_1000kb.RAWobserved, line 100):

```
8000000      13000000      1667.0
```

To normalize this entry using the KR normalization vector, one would divide 1667.0 by the 9th line ($(8000000/1000000)+1=9$) and the 14th line ($(13000000/1000000)+1=14$) of Rhesus-Macaque/1000kb_resolution_intrachromosomal/chr1/MQGE30/chrX_1000kb.KRnorm

The 9th line of the KR norm file is 0.980836359753368; the 14th line of the KR norm file is 0.9954328448063529. So the corresponding KR normalized entry for the entry above is $1667 / (0.980836359753368 * 0.9954328448063529) = 1707.36776057$. There will be some entries in the KR vector files which are NaNs – these correspond to rows which were too sparse and were removed before the normalization vector was computed (See Section II.b.4 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014). If the KR normalization vector file is completely empty or all NaNs, then the KR algorithm didn't converge on that particular matrix (likely due to sparsity of the matrix). In that case, one can use either the VC or SQRTVC normalizations or check a different resolution.

The four *expected files are expected vectors that can be used to construct O/E matrices that account for the increased number of contacts seen at short distances due to random polymer interactions driven by one-dimensional genome proximity (see the Glossary Appendix below and section II.c of the Extended Experimental Procedures of Rao & Huntley, et al., Cell, 2014 for more information about the construction of these expected vectors and O/E matrices). Note that these expected vectors only account for random polymer interactions driven by one-dimensional genome proximity and do not control for other features such as compartmentalization or contact domain structure. As such, they are not appropriate expected models when seeking to identify focal looping interactions. In identifying focal looping interactions, one must use local expecteds (see Section VI.a.4.i, Section VI.a.5.i, and Figure 3 of Rao & Huntley, et al., Cell 2014). These expected files are ordered such that first line of the expected vector file is the expected

number of contacts between two loci separated by 0*RES base pairs (0th diagonal of contact matrix), the second line of the file is the expected number of contacts between two loci separated by 1*RES base pairs (1st diagonal of contact matrix), and so on. So to create the O/E matrix, divide each entry $M_{i,j}$ by the expected value corresponding to the distance $j-i$.

For example, for the above entry of the Rhesus-Macaque 1000kb chr1 MAPQGE30 raw observed contact matrix, to get the corresponding entry of the O/E matrix divide by the 6th line ($((13000000-8000000)/1000000)+1=6$) of Rhesus-Macaque/1000kb_resolution_intrachromosomal/chr1/MQGE30/chrX_1000kb.RAWexpected. The 6th line of the raw expected file is 1240.2578. So the corresponding raw O/E entry for the entry above is $1667.0 / 1240.2578 = 1.3440754011$. To create KR, VC, or SQRTVC normalized O/E files, first construct the normalized observed matrices as above and then divide by the corresponding line of the normalized expected file (i.e. *.KRexpected, *.VCexpected, or *.SQRTVCexpected).

The interchromosomal contact matrix archives are found in GSE71831_[CELL_TYPE]_Interchromosomal.tar.gz, where [CELL_TYPE] denotes the cell type (e.g. Patski), and contain interchromosomal contact matrices at 8 different base pair delimited resolutions (1000 kb, 500kb, 250kb, 100kb, 50kb, 25kb, 10kb, 5kb). At each resolution, the files associated with each pair of chromosomes are provided in a separate subdirectory, chr[CHR1]_chr[CHR2]. Within each chromosome pair subdirectory, there are two further subdirectories (MAPQG0, MAPQGE30). The MAPQG0 subdirectory contains files associated with the contact matrices constructed from all read pairs that uniquely align to the genome (MAPQ>0). The MAPQGE30 subdirectory contains files associated with the contact matrices constructed from all read pairs that map to the genome with a MAPQ>=30. For example, the files associated with the contact matrix for chromosome 1 and chromosome 2 using MAPQ>=30 read pairs are in the subdirectory:

CELL_TYPE/100kb_resolution_interchromosomal/chr1_chr2/MAPQGE30/

The Patski interchromosomal chromosome pair folders have an additional subdirectory; besides for the MAPQG0 and MAPQGE30 subdirectories, they also contain a Phased subdirectory. The Phased directory further contains two subdirectories, maternal and paternal. All interchromosomal matrices in the phased folders are between two chromosomes of the same allele type. Thus for example Patski/1000kb_resolution_intrachromosomal/chr10_chr11/Phased/maternal/ contains the contact data for maternal chromosome 10 with maternal chromosome 11. Contact data for interactions between for example maternal chromosome 10 with paternal chromosome 11 is not available.

There are 7 text files associated with each interchromosomal contact matrix ([CHR1] represents the first chromosome number, [CHR2] represents the second chromosome number, and [RES] represents the resolution):

chr[CHR1]_chr[CHR2]_[RES]kb.RAWobserved
chr[CHR1]_[RES]kb.KRnorm
chr[CHR1]_[RES]kb.VCnorm
chr[CHR1]_[RES]kb.SQRTVCnorm
chr[CHR2]_[RES]kb.KRnorm

chr[CHR2]_[RES]kb.VCnorm
chr[CHR2]_[RES]kb.SQRTVCnorm

The *.RAWobserved files for the interchromosomal matrices are formatted the same way as the intrachromosomal matrices above except the i locus corresponds to CHR1 and the j locus corresponds to CHR2. One can create the normalized matrices as above except one should use the line corresponding to the i locus norm factor from the CHR1 normalization vector file and the line corresponding to the j locus norm factor from the CHR2 normalization vector file. (See the Glossary Appendix below and section II.b of the Extended Experimental Procedures of Rao & Huntley, et al., Cell, 2014 for more information about the different types of normalizations.)

Loop Annotation Files:

Filenames: GSE71831_[CELL_TYPE]_looplist.txt.gz

Cell types: RPE1-ΔDXZ4a, RPE1-ΔDXZ4i, RPE1-WT, Patski, AG08312, and GM12878

Under the main Series Record (GSE71831), there are files named GSE71831_[CELL_TYPE]_looplist.txt.gz, where CELL_TYPE represents each of the cell types analyzed in this study. These files contain Juicebox-loadable (www.aidenlab.org/juicebox) loop annotations returned by our loop calling algorithm, HiCCUPS. Annotations were performed using the 'hiccups' command with default settings from Juicebox Command Line Tools (<http://aidenlab.org/commandlinetools/>, see Durand, Shamim, et al., Cell Systems 2016 for more information).

These files contain a header line, followed by a line for every loop. There are 20 fields per line in the following format:

```
chromosome1  x1  x2  chromosome2  y1  y2  color  observed  expected_bottom_left  expected_donut  expected_horizontal  expected_vertical  fdr_bottom_left  fdr_donut  fdr_horizontal  fdr_vertical  number_collapsed  centroid1  centroid2  radius
```

Explanations of each field are as follows:

chromosome1 = chromosome2 = the chromosome that the loop is located on. HiCCUPS does not annotate enriched interactions between different types

x1,x2 = the coordinates of the upstream locus corresponding to the peak pixel (see the Experimental Procedures and VI.a.5.iv of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014 for a definition of the peak pixel)

y1,y2 = the coordinates of the downstream locus corresponding to the peak pixel (see the Experimental Procedures and VI.a.5.iv of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014 for a definition of the peak pixel)

color = the color that the feature will be rendered as if loaded in Juicebox (www.aidenlab.org/juicebox)

observed = the raw observed counts at the peak pixel (see the Experimental Procedures and VI.a.5.iv of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014 for a definition of the peak pixel)

expected_[bottom_left, donut, horizontal, vertical] = the expected counts calculated using the [bottom_left, donut, horizontal, vertical] filter (see Figure 3 and section VI.a.5.i of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014)

fdr_[bottom_left, donut, horizontal, vertical] = the q-value of the loop calculated using the [bottom_left, donut, horizontal, vertical] filter (see VI.a.5.ii of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014)

number_collapsed = the number of pixels that were clustered together as part of the loop call (see section VI.a.5.iv of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014)

centroid1 = the upstream coordinate of the centroid of the cluster of pixels corresponding to the loop (see section VI.a.5.iv of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014)

centroid2 = the downstream coordinate of the centroid of the cluster of pixels corresponding to the loop (see section VI.a.5.iv of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014)

radius = the Euclidean distance from the centroid of the cluster of pixels to the farthest pixel in the cluster of pixels (see section VI.a.5.iv of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014)

Contact Domain Annotation Files:

Filenames: GSE71831_[CELL_TYPE]_Arrowhead_domainlist_*_resolution.txt.gz

Cell types: RPE1-ΔDXZ4a, RPE1-ΔDXZ4i, RPE1-WT, Patski, AG08312, and GM12878

Under the main Series Record (GSE71831), there are files named GSE71831_[CELL_TYPE]_Arrowhead_domainlist_*_resolution.txt.gz where CELL_TYPE represents each of the cell types analyzed in this study and * gives the resolution at which the algorithm was performed. These files contain Juicebox-loadable (www.aidenlab.org/juicebox) domain annotations returned by our domain calling algorithm, Arrowhead (see Fig. 2, the Experimental Procedures, and Section IV.a. of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014). Annotations were performed using the 'arrowhead' command with default settings from Juicebox Command Line Tools (<http://aidenlab.org/commandlinetools/>, see Durand, Shamim, et al., Cell Systems 2016 for more information).

These files contain a header line, followed by a line for every domain. These files contain 12 fields per line in the following format:

```
chromosome1  x1  x2  chromosome2  y1  y2  color  corner_score  Uvar  Lvar
Usign  Lsign
```

Explanations of each field are as follows:

chromosome1 = chromosome2 = the chromosome that the domain is located on
x1,x2/y1,y2 = the interval spanned by the domain (contact domains manifest as squares on the diagonal of a Hi-C matrix and as such: x1=y1, x2=y2)

color = the color that the feature will be rendered as if loaded in Juicebox (www.aidenlab.org/juicebox)

corner_score = the corner score, a score indicating the likelihood that a pixel is at the corner of a contact domain. Higher values indicate a greater likelihood of being at the corner of a domain (see Section IV.a.3 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014)

Uvar = the variance of the upper triangle

Lvar = the variance of the lower triangle

Usign = -1*(sum of the sign of the entries in the upper triangle)
Lsign = sum of the sign of the entries in the lower triangle

Corner pixels which had both their left and right edges match to within 5 bins were merged to a single block call. In this case, the scores – given by the last 5 columns - are the mean of the all the corner pixels in the merged group, and the edges are given by the largest (in bp) values called for each edge. (See Section IV.a.3 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014 for details about these scores.)

GLOSSARY Appendix:

KRnorm = normalization vector obtained using the matrix balancing procedure of Knight and Ruiz on the intrachromosomal contact matrix, as reimplemented by us in Java. See Knight & Ruiz, IMA Journal of Numerical Analysis, 2012; section II.b.4 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell, 2014. Normalization factors are scaled such that the sum of the entries in the normalized contact matrix is equal to the sum of the entries in the raw contact matrix.

VCnorm = normalization vector obtained by calculating coverages (row-sums of the intrachromosomal contact matrix) for each locus as performed in Lieberman-Aiden, van Berkum, et al. See Lieberman-Aiden, van Berkum, et al., Science 2009; section II.b.1 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell, 2014. Note that these values are not the reciprocals of the row-sums as written in section II.b.1 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell 2014 but rather are proportional to the row sums themselves, hence why one must divide the raw observed matrix entries by the normalization factors rather than multiplying. Normalization factors are scaled such that the sum of the entries in the normalized contact matrix is equal to the sum of the entries in the raw contact matrix.

SQRTVCnorm = normalization vector obtained by taking the square root of the VCnorm vector. See section II.b.1 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell, 2014.

RAWexpected = a genome-wide model of interaction probability as a function of 1-dimensional genomic distance as introduced in Lieberman-Aiden, van Berkum, et al. and refined by us. See Lieberman-Aiden, van Berkum, et al., Science 2009; section II.c.1 of the Extended Experimental Procedures of Rao & Huntley, et al., Cell, 2014. Note that while the expected vector is calculated by taking averages over all matrix entries at a fixed distance genome-wide, the vectors are independently scaled for each chromosome such that the sum of the entries for a given intrachromosomal expected matrix is equal to the sum of the entries for the intrachromosomal observed matrix.

KRexpected = the same model of interaction probability as a function of 1-dimensional genomic distance as RAWexpected except calculated from the KR normalized contact matrix. Note that while the expected vector is calculated by taking averages over all matrix entries at a fixed distance genome-wide, the vectors are independently scaled for each chromosome such that the sum of the entries for a given intrachromosomal expected matrix is equal to the sum of the entries for the intrachromosomal observed matrix.

VCexpected = the same model of interaction probability as a function of 1-dimensional genomic distance as RAWexpected except calculated from the VC normalized contact matrix. Note that while the expected vector is calculated by taking averages over all matrix entries at a fixed distance genome-wide, the vectors are independently scaled for each chromosome such that the sum of the entries for a given intrachromosomal expected matrix is equal to the sum of the entries for the intrachromosomal observed matrix.

SQRTVCexpected = the same model of interaction probability as a function of 1-dimensional genomic distance as RAWexpected except calculated from the SQRTVC normalized contact matrix. Note that while the expected vector is calculated by taking averages over all matrix entries at a fixed distance genome-wide, the vectors are independently scaled for each chromosome such that the sum of the entries for a given intrachromosomal expected matrix is equal to the sum of the entries for the intrachromosomal observed matrix.